国际人工智能会议
AAAI 2021 论文北京预讲会

Regret Bounds for Online Kernel Selection in Continuous Kernel Space

Xiao Zhang, Shizhong Liao, Jun Xu*, Ji-Rong Wen

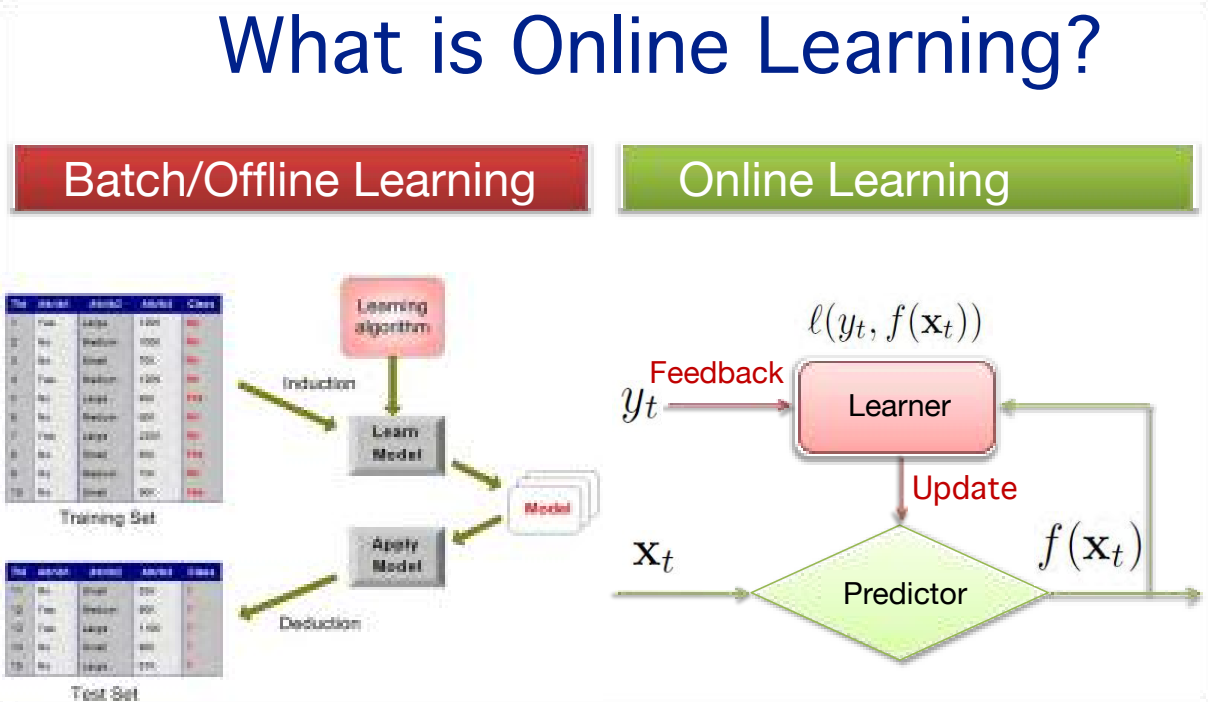
Gaoling School of Artificial Intelligence, Renmin University of China

College of Intelligence and Computing, Tianjin University

zhangx89@ruc.edu.cn

Outline

Offline / Online Learning



Offline / Online Model Selection

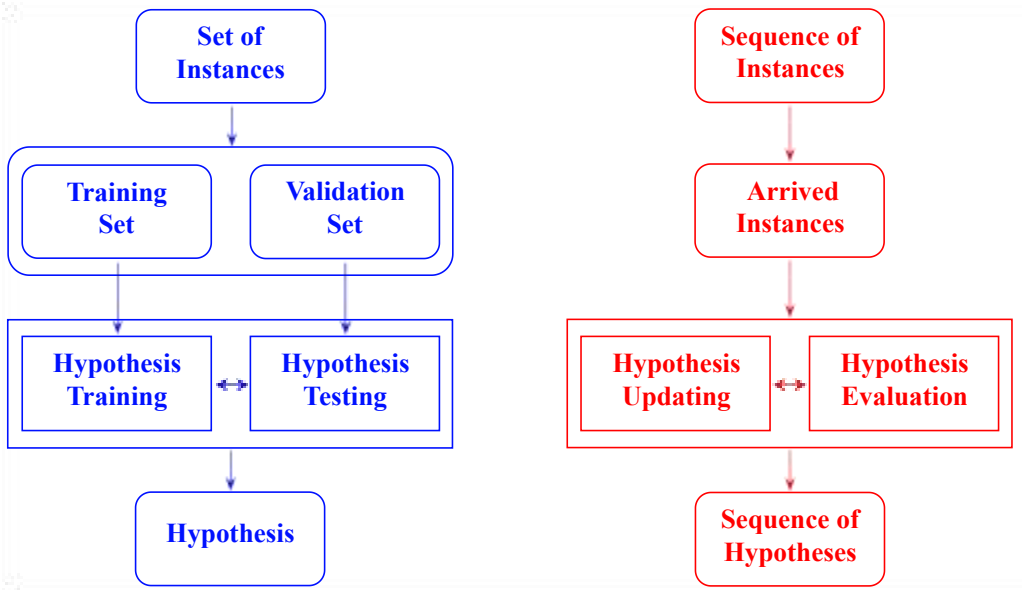
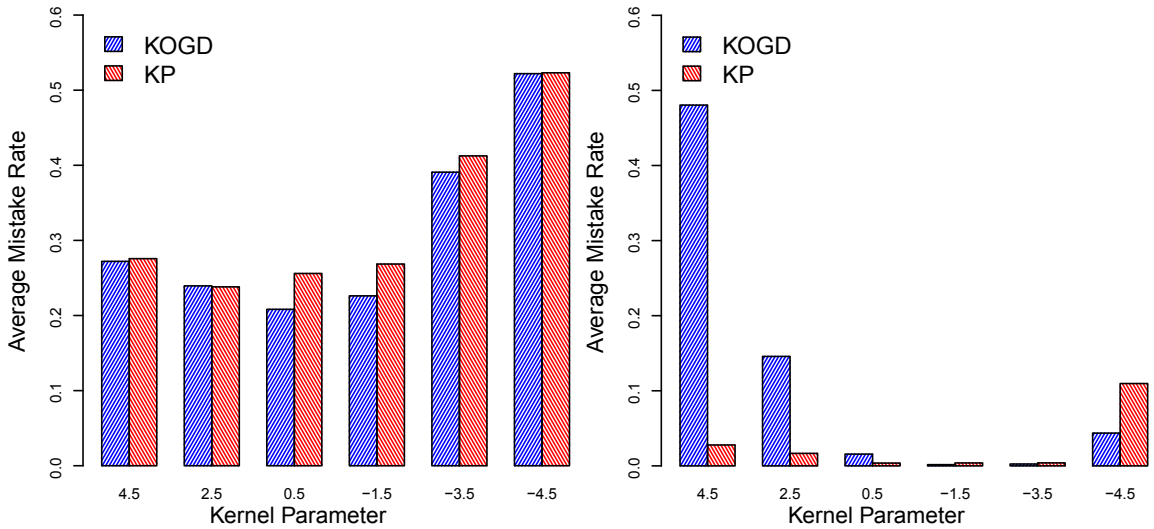


Figure 1: Comparison between offline model selection and online model selection

Necessity of Online Model Selection



(a) `spambase` (b) `mushrooms`

Figure 2: Performances of Kernelized Online Gradient Decent (KOGD) and Kernel Perceptron (KP) using different Gaussian kernels for online classification

Candidate Kernels

- **Finite kernel set**

  containing a finite number of candidate kernels:

  $$\widehat{\mathcal{H}}_N = \{\mathcal{H}_i\}_{i\in[N]},$$

  where N is the number of candidate kernels.

- **Continuous kernel space**

  containing continuously many candidate kernels:

  $$\widehat{\mathcal{H}}_\Omega = \{\mathcal{H}_\sigma \mid \sigma \in \Omega\},$$

  where Ω is the parameter interval of candidate kernels，

Regret of Online Kernel Selection

Definition 1 (Worse-Case Regret of Online Kernel Selection)

Given a continuous kernel space $\widehat{\mathcal{H}}_\Omega$
Assuming that $\{f_t\}_{t\in[T]} \subseteq \widehat{\mathcal{H}}_\Omega$ is a hypothesis sequence generated by an online kernel selection procedure after T rounds,
the *worse-case regret* (regret) of online kernel selection is defined as follows:

$$\mathcal{R}_{\text{reg}}\left(\{f_t\}_{t\in[T]}, g\right) := \sum_{t=1}^{T} \left[\ell\left(f_t(\boldsymbol{x}_t), y_t\right) - \ell\left(g(\boldsymbol{x}_t), y_t\right)\right] = o(T),$$

where $g \in \widehat{\mathcal{H}}_\Omega$ is the competing hypothesis,
which is typically defined as the best hypothesis in hindsight:

$$g = f^* := \arg\min_{f\in\widehat{\mathcal{H}}_\Omega} \sum_{t=1}^{T} \ell\left(f(\boldsymbol{x}_t), y_t\right).$$

New Challenges of Online Kernel Selection

(1) High computational complexities

- Quadratic time complexity [Singh and Príncipe, 2011]
- Linear space complexity [Chen et al., 2016]

(2) Lack of theoretical guarantees

- Regret bounds dependent on the number of kernels [Yang et al., 2012]
- Expected regret bounds in continuous kernel space [Zhang and Liao, 2020]

Main Contribution

Table 1: Comparison with the existing online kernel selection approaches

	Computational complexities			**Theoretical guarantees**	
	Time (per round)	Time (overall)	Space	#Candidate	Regret bound
Existing	Linear	Quadratic	Linear	Finite	Square root
Proposed	Constant	Linear	Constant	Infinite	Square root

Learning Framework

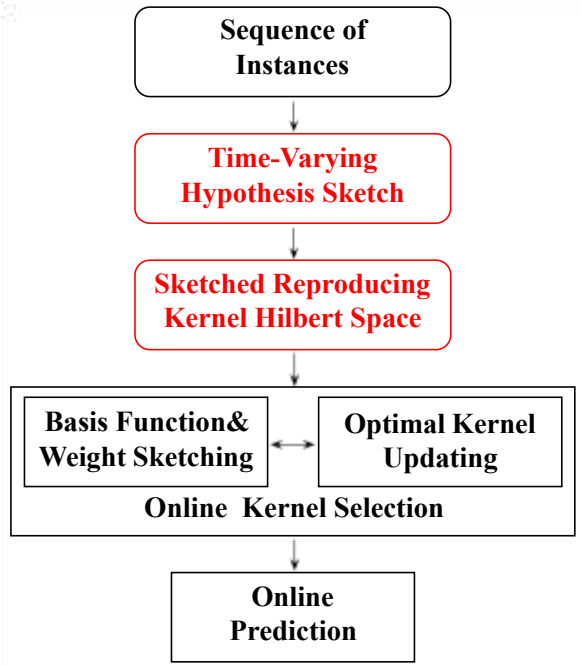


Figure 3: Online kernel selection using time-varying hypothesis sketching

Time-Varying Hypothesis Sketch

- Time-varying hypothesis sketch at round t

$$f_{\sigma_t,t}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}^{(t)}_{\sigma_t}(\cdot) \rangle, \quad \sigma_t \in \Omega,$$

  where $\mathcal{V}_t = \{\tilde{\boldsymbol{x}}_i\}_{i=1}^{|\mathcal{V}_t|} \subseteq \mathcal{X}$ is a buffer.

- Basis vector and its corresponding weight vector

$$\boldsymbol{\psi}^{(t)}_{\sigma_t}(\cdot) = \left[\kappa_{\sigma_t}(\cdot, \tilde{\boldsymbol{x}}_1), \ldots, \kappa_{\sigma_t}(\cdot, \tilde{\boldsymbol{x}}_{|\mathcal{V}_t|})\right]^\intercal, \quad \tilde{\boldsymbol{x}}_i \in \mathcal{V}_t,$$
$$\boldsymbol{\omega}^{(t)} = \left[\omega_1^{(t)}, \omega_2^{(t)}, \ldots, \omega_{|\mathcal{V}_t|}^{(t)}\right]^\intercal \in \mathbb{R}^{|\mathcal{V}_t|}.$$

Learning Framework

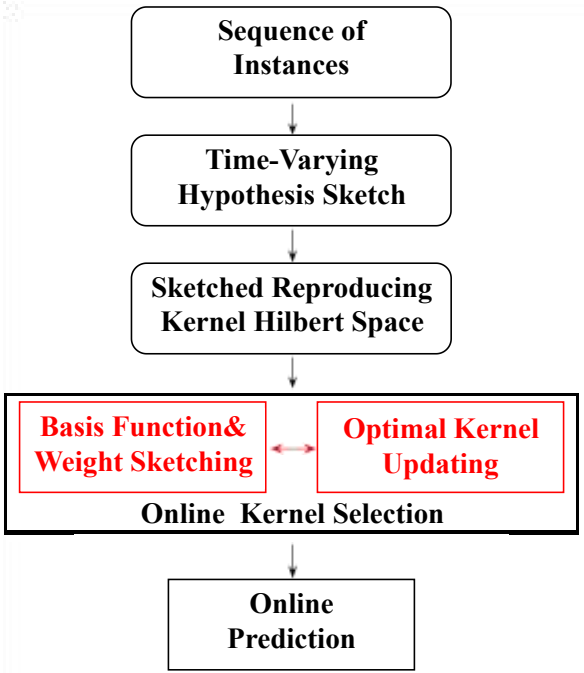


Figure 4: Online kernel selection using time-varying hypothesis sketching

Two Online Kernel Selection Categories

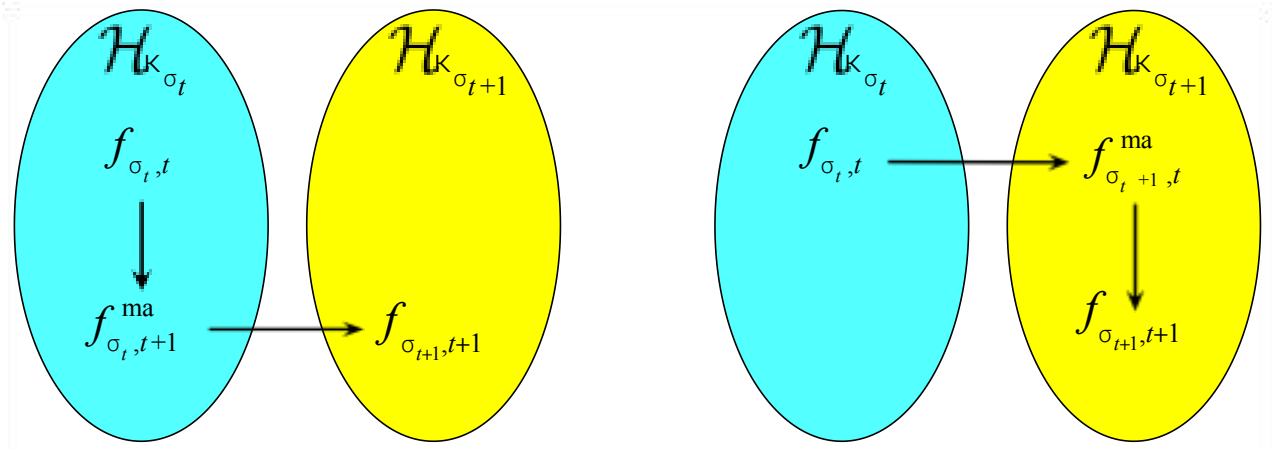


Figure 5: Two Online Kernel Selection Categories at round t: OKS-SPT (left) denotes the online kernel selection by **selection-post-training**; OKS-TPS (right) denotes the online kernel selection by **training-post-selection**.

Two Online Kernel Selection Categories

Category 1: Online Kernel Selection by Selection-Post-Training (OKS-SPT)

Require: the continuous kernel space $\mathcal{K}_\Omega$, initial kernel κ_{σ_1}

1: Initialize the weight vector $\boldsymbol{\omega}^{(1)} = \mathbf{0}$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute the hypothesis sketch $f_{\sigma_t,t}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t)}(\cdot) \rangle$
4: Predict $\hat{y}_t = \mathrm{sgn}(f_{\sigma_t,t}(\boldsymbol{x}_t))$ for classification or $\hat{y}_t = f_{\sigma_t,t}(\boldsymbol{x}_t)$ for regression
5: Maintain the buffer $\mathcal{V}_{t+1} = \textsc{BufferMaintenance}(\mathcal{V}_t, \boldsymbol{z}_t)$ and obtain $f_{\sigma_t,t}^{\mathrm{ma}} = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t+1)}(\cdot) \rangle$
6: Update the weight vector $\boldsymbol{\omega}^{(t+1)} = \textsc{WeightUpdating}(f_{\sigma_t,t}^{\mathrm{ma}}, \boldsymbol{z}_t)$ and obtain $f_{\sigma_t,t+1}^{\mathrm{ma}} = \langle \boldsymbol{\omega}^{(t+1)}, \boldsymbol{\psi}_{\sigma_t}^{(t+1)}(\cdot) \rangle$
7: **if** the buffer changes **then**
8: Select $\kappa_{\sigma_{t+1}} = \textsc{KernelSelection}(\kappa_{\sigma_t}, \boldsymbol{z}_t)$ from $\mathcal{K}_\Omega$ and obtain $f_{\sigma_{t+1},t+1}$
9: **else**
10: $\kappa_{\sigma_{t+1}} = \kappa_{\sigma_t}$
11: **end if**
12: **end for**

Two Online Kernel Selection Categories

Category 2: Online Kernel Selection by Training-Post-Selection (OKS-TPS)

Require: the continuous kernel space $\mathcal{K}_\Omega$, initial kernel κ_{σ_1}

1: Initialize the weight vector $\boldsymbol{\omega}^{(1)} = \mathbf{0}$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute the hypothesis sketch $f_{\sigma_t,t}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}^{(t)}_{\sigma_t}(\cdot) \rangle$
4: Predict $\hat{y}_t = \mathrm{sgn}(f_{\sigma_t,t}(\boldsymbol{x}_t))$ for classification or $\hat{y}_t = f_{\sigma_t,t}(\boldsymbol{x}_t)$ for regression
5: Select a kernel $\kappa_{\sigma_{t+1}} = \textsc{KernelSelection}(\kappa_{\sigma_t}, \boldsymbol{z}_t)$ from $\mathcal{K}_\Omega$
6: Maintain the buffer $\mathcal{V}_{t+1} = \textsc{BufferMaintenance}(\mathcal{V}_t, \boldsymbol{z}_t)$ and obtain $f^{\mathrm{ma}}_{\sigma_{t+1},t} = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}^{(t+1)}_{\sigma_{t+1}}(\cdot) \rangle$
7: Update the weight vector $\boldsymbol{\omega}^{(t+1)} = \textsc{WeightUpdating}(f^{\mathrm{ma}}_{\sigma_{t+1},t}, \boldsymbol{z}_t)$ and obtain $f_{\sigma_{t+1},t+1}$
8: **end for**

Our Theory: Regret Analysis

Theorem 2 (Regret of OKS-SPT)

Let $\mathcal{K}_\Omega$ be a continuous kernel space that contains Gaussian kernel functions, $\{f_{\sigma_t,t}\}_{t=1}^T \subseteq \widehat{\mathcal{H}}$ be the hypothesis sketch sequence generated by OKS-SPT.
Define $\widehat{D}(f,g) = \max_{t\in[T]} |f(\boldsymbol{x}_t) - g(\boldsymbol{x}_t)|$, where $f, g \in \widehat{\mathcal{H}}$.
Assume $C_{\max} = \max_{i,j\in[T]} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ and $R = \sup_{f\in\widehat{\mathcal{H}}} \|f\|_{\widehat{\mathcal{H}}}$, then there exists a constant $C > 0$ such that

$$\widehat{\mathrm{Reg}}_T(\{f_{\sigma_t,t}\}_{t=1}^T, \bar{f}^*) \leq \underbrace{U_1}_{\text{Optimization}} + \underbrace{U_2}_{\text{Estimation}} + \underbrace{U_3}_{\text{Approximation}},$$

where $U_2 = \widehat{D}(\bar{f}^*_{\sigma_t}, f^*_{\sigma_t})$,

$$U_1 = 2RC\sqrt{\mu}\,[T - O(\ln T)] + \frac{R^2}{2\eta_f} + \frac{\left(C\sqrt{\mu}+1\right)^2}{2}\eta_f T,$$

$$U_3 = 2C_{\max}\frac{\sigma_{\max}}{\sigma_{\min}^3}\widehat{D}(\bar{f}^*_{\sigma_t}, f_{\sigma_t,t+1}) + \frac{(\bar{\sigma}^*)^2}{2\eta_\sigma} + \frac{\left[\sigma_{\min}^{-3}C_{\max}\widehat{D}(\bar{f}^*_{\sigma_t}, f_{\sigma_t,t+1}) + L\right]^2}{2}\eta_\sigma T.$$

Our Theory: Regret Analysis

(1) **Selection-post-training (OKS-SPT)**: $O(\sqrt{T})$ 后悔界

- Stepsize: $\eta_f, \eta_\sigma = O(1/\sqrt{T})$
- Threshold value for buffer: $\nu = O(1/T)$
- Hypothesis value: $O(1/\sqrt{T})$

(2) **Selection-post-training (OKS-TPS)**: $O(\sqrt{T})$ 后悔界

- Stepsize: $\eta_f, \eta_\sigma = O(1/\sqrt{T})$
- Threshold value for buffer: $\nu = O(1/T)$
- Conditions for kernel matrix:

$$\det \boldsymbol{K}_{\bar{\sigma}^*, \mathcal{V}_t \cup \{\boldsymbol{x}_t\}} / \det \boldsymbol{K}_{\bar{\sigma}^*, \mathcal{V}_t} = O(\mu),$$

$$\left\| \left(\boldsymbol{K}_{\bar{\sigma}^*, \mathcal{V}_t}\right)^\dagger \boldsymbol{\psi}^{(t)}_{\bar{\sigma}^*}(\boldsymbol{x}_t) - \left(\boldsymbol{K}_{\sigma_{t+1}, \mathcal{V}_t}\right)^\dagger \boldsymbol{\psi}^{(t)}_{\sigma_{t+1}}(\boldsymbol{x}_t) \right\| = O(\mu).$$

Our Theory: Comparison of Theoretical Results

Table 2: Comparison among online kernel selection approaches

Approach	Computational Complexities			Regret Guarantees	
	#Updates	Time (overall)	Space	Candidate	Regret bound
OKS	T	$O(T^2 + NT)$	$O(T)$	Finite	$O\left(\sqrt{N(\ln N)T}\right)$
MS-FTPL	T	$O\left(NT^2\right)$	$O(NT)$	Finite	$O\left(\sqrt{NT \ln T}\right)$
OKL-GD	T	$O(T^2)$	$O(T)$	Continuous	–
RRF	T	$O(T^2)$	$O(T)$	Continuous	–
OKS-SPT	$\ln(T)$	$O((\ln T)^2 T)$	$O((\ln T)^2)$	Continuous	$O\left(\sqrt{T}\right)$
OKS-TPS	T	$O((\ln T)^2 T)$	$O((\ln T)^2)$	Continuous	$O\left(\sqrt{T}\right)$

- T: the number of rounds.
- N: the number of candidate kernels.

Main Results

Empirical Verification

Table 3: Performances of OKL-GD, OKS, RRF and the proposed OKS-SPT, OKS-TPS for online classification w.r.t. the mistake rate (%) and the running time (s).

Algorithm	german		spambase		mushrooms	
	Mistake rate	Time	Mistake rate	Time	Mistake rate	Time
OKL-GD	34.960 ± 1.518	**0.194**	36.031 ± 0.421	6.700	4.226 ± 0.910	37.230
OKS	42.320 ± 1.307	0.226	34.355 ± 0.372	4.083	9.441 ± 0.282	9.787
RRF	31.140 ± 0.114	0.372	44.961 ± 0.820	4.767	16.166 ± 0.964	21.750
OKS-SPT	29.920 ± 0.286	0.244	**28.436 ± 0.213**	**2.533**	6.585 ± 0.246	**4.240**
OKS-TPS	**29.760 ± 0.270**	0.296	28.450 ± 0.188	2.590	**3.139 ± 0.481**	6.910
Algorithm	a9a		w7a		ijcnn1	
	Mistake rate	Time	Mistake rate	Time	Mistake rate	Time
OKL-GD	23.936 ± 0.008	321.525	2.975 ± 0.062	857.268	9.575 ± 0.012	134.880
OKS	23.617 ± 0.127	1053.420	7.637 ± 0.024	943.855	9.578 ± 0.184	618.520
RRF	23.931 ± 0.001	152.265	2.978 ± 0.004	674.735	9.574 ± 0.001	39.290
OKS-SPT	**20.368 ± 0.659**	**39.360**	2.675 ± 0.023	**94.530**	9.478 ± 0.003	**33.860**
OKS-TPS	22.379 ± 0.192	48.815	**2.631 ± 0.010**	96.395	**9.440 ± 0.002**	35.165

Conclusion

- Two online kernel selection categories via time-varying hypothesis sketching.
- Meet the new challenges of online kernel selection.
- A time-varying sketching approach to online model selection.

Main References

[Chen et al., 2016] Chen, B., Liang, J., Zheng, N., and Príncipe, J. C. (2016).
Kernel least mean square with adaptive kernel size.
Neurocomputing, 191:95–106.

[Singh and Príncipe, 2011] Singh, A. and Príncipe, J. C. (2011).
Information theoretic learning with adaptive kernels.
IEEE Transactions on Signal Processing, 91(2):203–213.

[Yang et al., 2012] Yang, T., Mahdavi, M., Jin, R., Yi, J., and Hoi, S. C. (2012).
Online kernel selection: Algorithms and evaluations.
In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 22–26.

[Zhang and Liao, 2020] Zhang, X. and Liao, S. (2020).
Hypothesis sketching for online kernel selection in continuous kernel space.
In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 2498–2504.

国际人工智能会议
AAAI 2021 论文北京预讲会

THANKS

2020.12.19